Neighborhood comparison between Madrid and Barcelona

14th June 2020

Author: Claudio Pravia Silvera

1. Introduction

The business problem we want to address is to provide franchise operators with information that makes it easier to identify the probability of success when opening a franchise in either Madrid or Barcelona.

To do this we will compare the two cities to discover the neighborhoods for similarities related to the category and number of venues that they have. We will cluster the neighborhoods using k-means and will then examine the clusters to find similarities between the two cities. We will analyse what makes them more attractive to tourists.

The target audience is therefore the franchise operators that focus on tourism in these two cities and needs to know where to open the franchise.

2. Data acquisition and cleaning

First, we will extract the list of neighborhoods of the two cities. After that, we will use Nominatim to obtain the geographical coordinates for the neighborhoods by using the neighborhoods' name, and finally using Foursquare to obtain the different venues that exist in each of the neighborhoods.

2.1 Data sources

The list of neighborhoods' data was obtained using the following Wikipedia links:

https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid
The neighborhood names are the column "Nombre".

https://es.wikipedia.org/wiki/Distritos_de_Barcelona
The neighborhood names are the column "Barrios" that are a set that will be separated.

We will use a 'get' request using the Foursquare API, to obtain the following venue data: name, latitude, longitude and category.

2.1.1 Madrid.

The first step was to get the Names of the neighborhoods.

```
[2]: url = 'https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid'
     html = requests.get(url).content
     df_list = pd.read_html(html)
     df_mad = df_list[-1]
     df_mad.head()
```

[2]:

	Distrito	Número	Nombre	Superficie (km²)[2]	Imagen
0	Centro	11	Palacio	1,471 km²	NaN
1	Centro	12	Embajadores	1,032 km²	NaN
2	Centro	13	Cortes	0,592 km²	NaN
3	Centro	14	Justicia	0,742 km²	NaN
4	Centro	15	Universidad	0,947 km²	NaN

2.1.2 Barcelona.

In this case the Names of the neighborhoods came as a set in one field by district and though they were labeled with a number, some of them were separated by commas and other by "and": sometimes the "and" was in Spanish ("y") and other times in Catalan ("i"). The decision was made to save them to a csv file and do the cleansing manually.

```
url = 'https://es.wikipedia.org/wiki/Distritos_de_Barcelona'
html = requests.get(url).content
df_list = pd.read_html(html)
df_bcn = df_list[-2]
df_bcn = df_bcn.rename(columns={'Barrios (nº)': 'Neighborhood'})
df_bcn2= df_bcn['Neighborhood']
df_bcn = df_bcn2.to_frame()
df_bcn.head()
```

	Neighborhood
0	El Raval (1), Barrio Gótico (2), La Barcelonet...
1	El Fort Pienc (5), Sagrada Familia (6), Dreta ...
2	Pueblo Seco (11), La Marina del Prat Vermell (...
3	Les Corts (19), La Maternidad y San Ramón (20)...
4	Vallvidrera, el Tibidabo i les Planes (22), Sa...

```
df_bcn2 = pd.read_csv('BcnBarrios1.csv', encoding="latin-1")
df_barcelona = df_bcn2.T.reset_index()
df_barcelona.columns = ['Neighborhood']
df_barcelona.head()
```

	Neighborhood
0	El Raval
1	Barrio Gótico
2	La Barceloneta
3	Santa Caterina i la Ribera
4	El Fort Pienc

2.2 Data cleansing

When getting the coordinates, an exception was programmed to discard the neighborhoods without data.

2.2.1 Madrid.

We got the coordinates using Nominatim:

```
df_madll.head()
```

	Neighborhood	Latitude	Longitude
0	Palacio	40.415129	-3.715618
1	Embajadores	40.409681	-3.701644
2	Cortes	40.414779	-3.697584
3	Justicia	40.423957	-3.695747
4	Universidad	40.425409	-3.705989

The list below shows the neighborhoods names for which we could not get the coordinates.

```
data = []
df_madll = pd.DataFrame(data, columns=['Neighborhood', 'Latitude', 'Longitude'])

for i in range(len(df_madrid)) :
    address = df_madrid.iloc[i,0]+', Madrid, Spain'
    try:
        geolocator = Nominatim(user_agent="ny_explorer")
        location = geolocator.geocode(address)
        latitude = location.latitude
        longitude = location.longitude

        df_madll = df_madll.append({'Neighborhood':df_madrid.iloc[i,0],'Latitude':latitude,'Longitude':longitude}, ignore_index=True)

    except:
        print( df_madrid.iloc[i,0])
```

```
Los Cármenes
Valderrivas
Casco Histórico de Barajas
```

2.2.2 Barcelona.

This is the list with coordinates for Barcelona.

```
df_bcnll.head()
```

	Neighborhood	Latitude	Longitude
0	El Raval	41.379518	2.168368
1	Barrio Gótico	41.383395	2.176912
2	La Barceloneta	41.380653	2.189927
3	Santa Caterina i la Ribera	41.386650	2.184194
4	El Fort Pienc	41.395925	2.182325

The list below shows the neighborhoods names for which we could not get the coordinates.

```
data = []
df_bcnll = pd.DataFrame(data, columns=['Neighborhood', 'Latitude', 'Longitude'])

for i in range(len(df_barcelona)) :
    address = df_barcelona.iloc[i,0]+', Barcelona, Spain'
    try:
        geolocator = Nominatim(user_agent="ny_explorer")
        location = geolocator.geocode(address)
        latitude = location.latitude
        longitude = location.longitude

        df_bcnll = df_bcnll.append({'Neighborhood':df_barcelona.iloc[i,0],'Latitude':latitude,'Longitude':longitude}, ignore_index=True)

    except:
        print( df_barcelona.iloc[i,0])
```

```
La Antigua Izquierda del Ensanche
La Nueva Izquierda del Ensanche
Pueblo Seco
Vallcarca y los Penitentes
Villa de Gracia
El Valle de Hebrón
El Campo del Arpa del Clot
Pueblo Nuevo
San Martín de Provensals
```

2.3 Feature selection

Using Foursquare data, we identified the top 50 venues for each neighborhood that are within a 200-meter radius from the center of the neighborhood. We encoded and normalized the venues per neighborhood.

2.3.1 Madrid.

	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bar	Basketball Court	Bed & Breakfast	Beer Bar	Beer Garden	Beer Store	Big Box Stor
0	Abrantes	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.200000	0.000000	0.000000	0.0	0.000000	0.000000	0.00000	0.00000
1	Acacias	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.030303	0.060606	0.000000	0.0	0.000000	0.000000	0.00000	0.00000
2	Adelfas	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.058824	0.000000	0.000000	0.058824	0.000000	0.000000	0.0	0.000000	0.000000	0.00000	0.00000
3	Alameda de Osuna	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.111111	0.000000	0.000000	0.0	0.000000	0.000000	0.00000	0.00000
4	Almagro	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.026316	0.000000	0.0	0.026316	0.000000	0.00000	0.00000
5	Almenara	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.285714	0.000000	0.0	0.000000	0.000000	0.00000	0.00000
6	Almendrales	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.00000	0.00000
7	Aluche	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.125000	0.000000	0.0	0.000000	0.000000	0.00000	0.00000
8	Amposta	0.00	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.250000	0.250000	0.000000	0.0	0.000000	0.000000	0.00000	0.00000

2.3.2 Barcelona.

	Neighborhood	Yoga Studio	Accessories Store	African Restaurant	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	Auto Workshop	BBQ Joint	Baby Store	Bakery	Bar	Basketball Court	Basketball Stadium	Beach	Beach Bar
0	Baix Guinardó	0.00000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Barrio Gótico	0.00000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.02	0.00	0.000000	0.000	0.000000	0.000000	0.00	0.02	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Camp d'en Grassot i Gràcia Nova	0.00000	0.000000	0.000000	0.00	0.00	0.00	0.025641	0.00	0.00	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.076923	0.076923	0.025641	0.000000	0.000000	0.000000
3	Can Baró	0.00000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.142857	0.000000	0.000000	0.000000
4	Can Peguera	0.00000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	Canyelles	0.00000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.250000	0.000000	0.000000	0.000000	0.000000	0.000000
6	Ciudad Meridiana	0.00000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	Congrés i els Indians	0.00000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	Diagonal Mar i Front Marítim del Poblenou	0.00000	0.000000	0.000000	0.00	0.00	0.00	0.000000	0.00	0.00	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.033333	0.000000	0.000000	0.033333	0.066667
9	Dreta de l''Eixample	0.00000	0.000000	0.000000	0.00	0.00	0.00	0.020000	0.00	0.00	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.020000	0.000000	0.000000	0.000000	0.000000	0.000000

We got the 5 most common venues for each neighborhood.

2.3.3 Madrid venues.

```
neighborhoods_venues_sorted.head()
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Abrantes	Metro Station	Pizza Place	Grocery Store	Bakery	Fast Food Restaurant
1	Acacias	Spanish Restaurant	Café	Pizza Place	Restaurant	Supermarket
2	Adelfas	Spanish Restaurant	Breakfast Spot	Bakery	Diner	Coffee Shop
3	Alameda de Osuna	Tapas Restaurant	Italian Restaurant	Metro Station	Pizza Place	Smoke Shop
4	Almagro	Spanish Restaurant	Restaurant	Pub	French Restaurant	Salad Place

2.3.4 Barcelona venues.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Baix Guinardó	Tapas Restaurant	Supermarket	Latin American Restaurant	Falafel Restaurant	School
1	Barrio Gótico	Plaza	Spanish Restaurant	Tapas Restaurant	Wine Bar	Hotel
2	Camp d'en Grassot i Gràcia Nova	Tapas Restaurant	Diner	Bar	Bakery	Café
3	Can Baró	Grocery Store	Chinese Restaurant	Plaza	Spanish Restaurant	Basketball Court
4	Can Peguera	Park	Wings Joint	Cupcake Shop	Farmers Market	Falafel Restaurant

3. Methodology

We wanted to classify the different neighborhoods by the amount of common venues that they have.

3.1 Machine learning used

We used K-Means clustering to cluster the neighborhoods in different groups to find patterns that can help us get the findings that will help to solve our business problem.

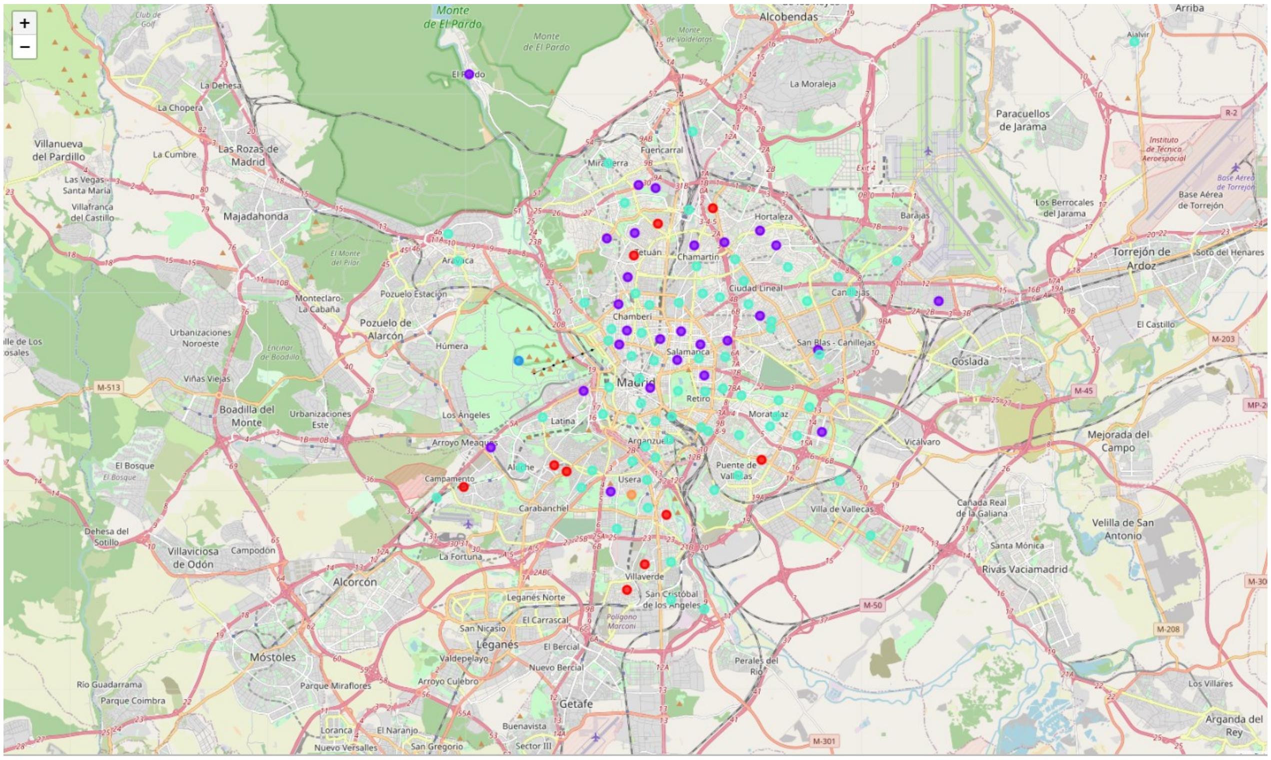
3.2 Exploratory Data Analysis

We used different values of k and found k=6 to be the optimal. Less than that will confound the clusters and more will include clusters that are not relevant.
We also changed the radius of the circles we used to identify venues in each neighborhood. We started with 500 meters, and then reduced to 300 meters and finally to 200 meters. We realized that with the larger radius some venues will appear to be part of more than one neighborhood and that was to be avoided.

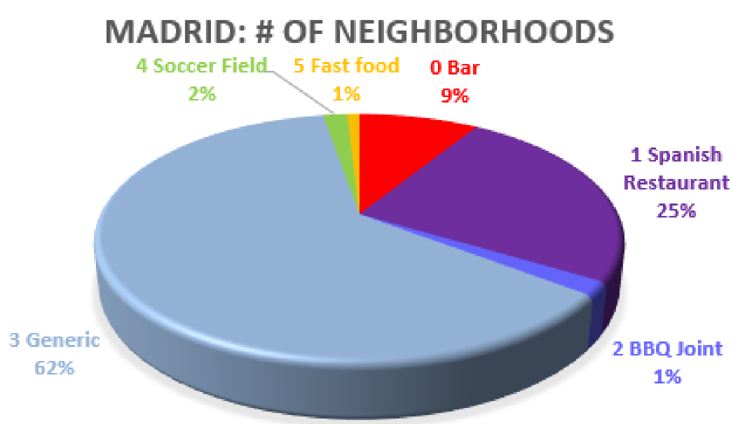
4. Results

4.1 Clustering neighborhoods in Madrid.

The following figure shows a map of Madrid with the neighborhoods shown in different colors for each cluster.

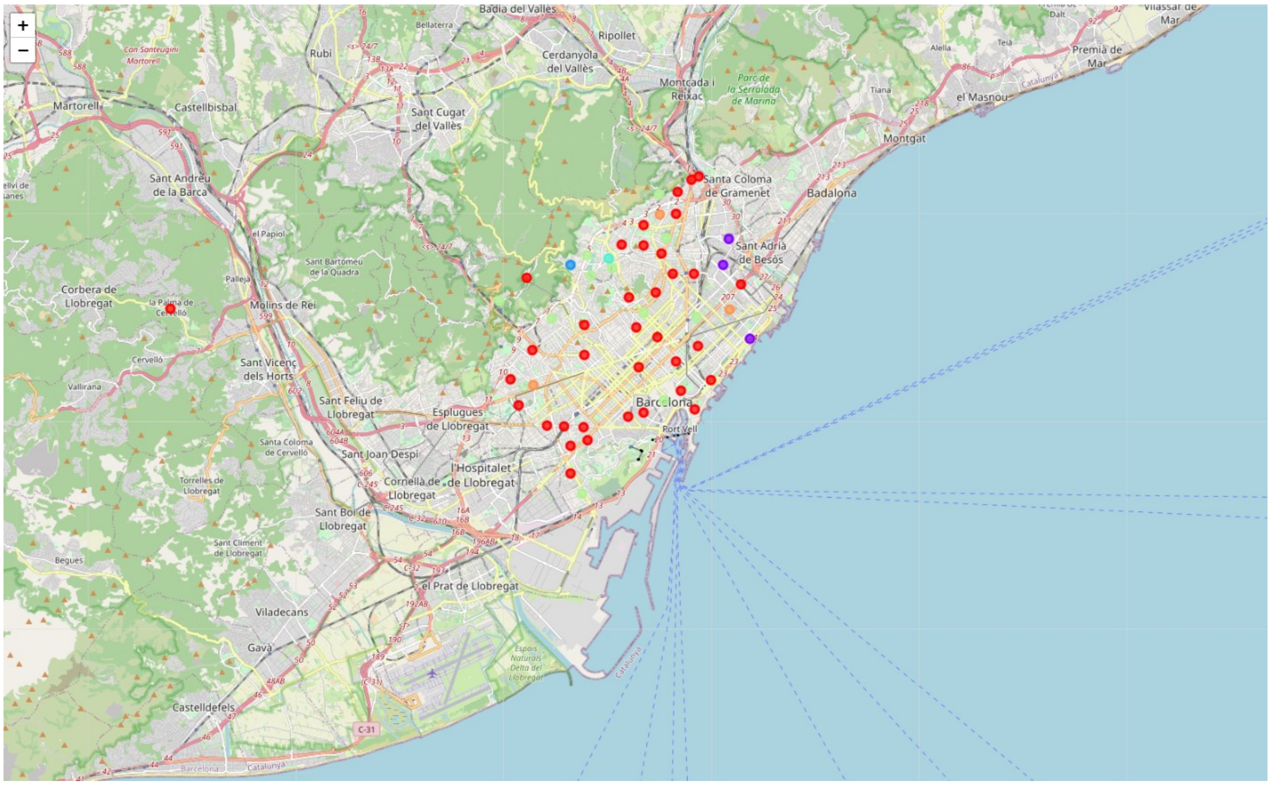


This one shows a chart with the percentage of the number of neighborhoods that fall into each cluster in Madrid.

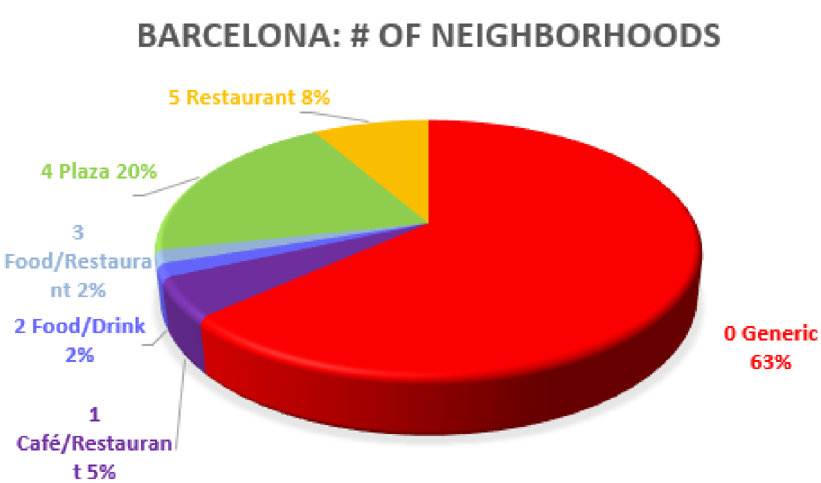


4.2 Clustering neighborhoods in Barcelona

This figure shows a map of Barcelona with the neighborhoods shown in different colors for each cluster.



And this one shows a chart with the percentage of the number of neighborhoods that fall into each cluster for Barcelona.



5. Discussion

We see in both cases a predominant cluster of more than 60% that contains venues that we labeled as Generic. The venues in these clusters have many different categories that cannot be summarized. In Madrid, the second largest cluster has a clear bias for the category “Spanish Restaurant” while the third has the category “Bar”.
In the case of Barcelona, the second largest cluster has the category “Plaza” (meaning “park”) as predominant while the third has the category “Restaurant”.

6. Conclusion

We can recommend, for a franchise in Spanish food that would like to open in Madrid, choose one of the neighborhoods in the Madrid cluster 1.
In Barcelona, that is in Catalonia, a province that wants to be independent, there is no cluster related to “Spanish Restaurant”. The Barcelona cluster 5 though, shows neighborhoods with the main category as “Restaurants”. We can therefore recommend a more general food franchise to be opened in these neighborhoods.

7. Future directions

We could deepen this study by concentrating only in restaurants and then try to find more specific relationships between the type of cuisine and the neighborhoods in Madrid and Barcelona.